A Many-to-Many Deep Learning Imaging Decoder for Real-Time Retrieval of the Human Pose Behind Visual Obstruction

Full paper on arXiv
Published to IEEE ICMLA 2019

Kevin Meng
Plano West Senior High School, Plano, Texas

Introduction

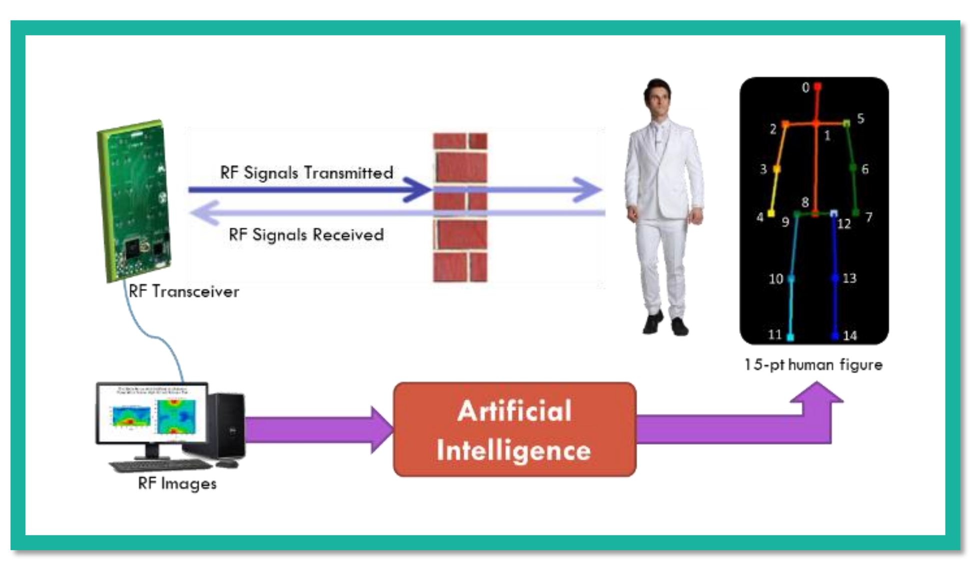
Being able to see through walls has always been a dream of mankind. Especially in search & rescue, non-invasive healthcare, and military situations, through-wall vision could prove instrumental to saving lives.

Although visible light cannot penetrate walls, Radio-Frequency (RF) signals travel through many kinds of obstruction and reflect off humans.

But using RF reflections for imaging also has its tradeoffs. Rule-based algorithms struggle to generalize to complex reflection patterns and different environments. Existing machine-learning works also struggle to explicitly address several important physical challenges of RF.

Project Goal

This project proposes a novel many-to-many deep-learning imaging architecture that converts a continuous stream of RF images into a human figure while solving several key physical limitations of the wave.



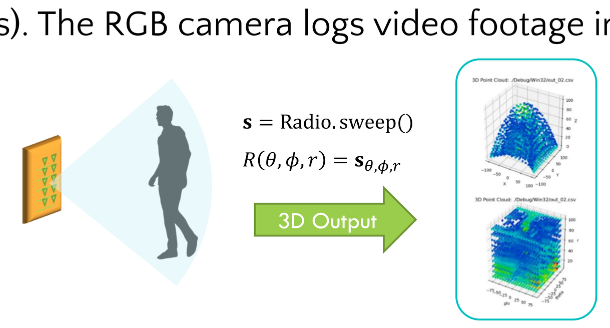
Sensors (I)

1) A Frequency Modulated Continuous Wave (FMCW) Antenna Array that transceives RF signals at frequency $[3.3GHz - 10GHz]$. Transmits at 1/1000 the power of Wi-Fi. Uses Walabot Developer SDK.

2) An RGB webcam is co-located with the RF antenna to collect color-channel video in parallel.

Data Acquisition & Pre-Processing (II/III)

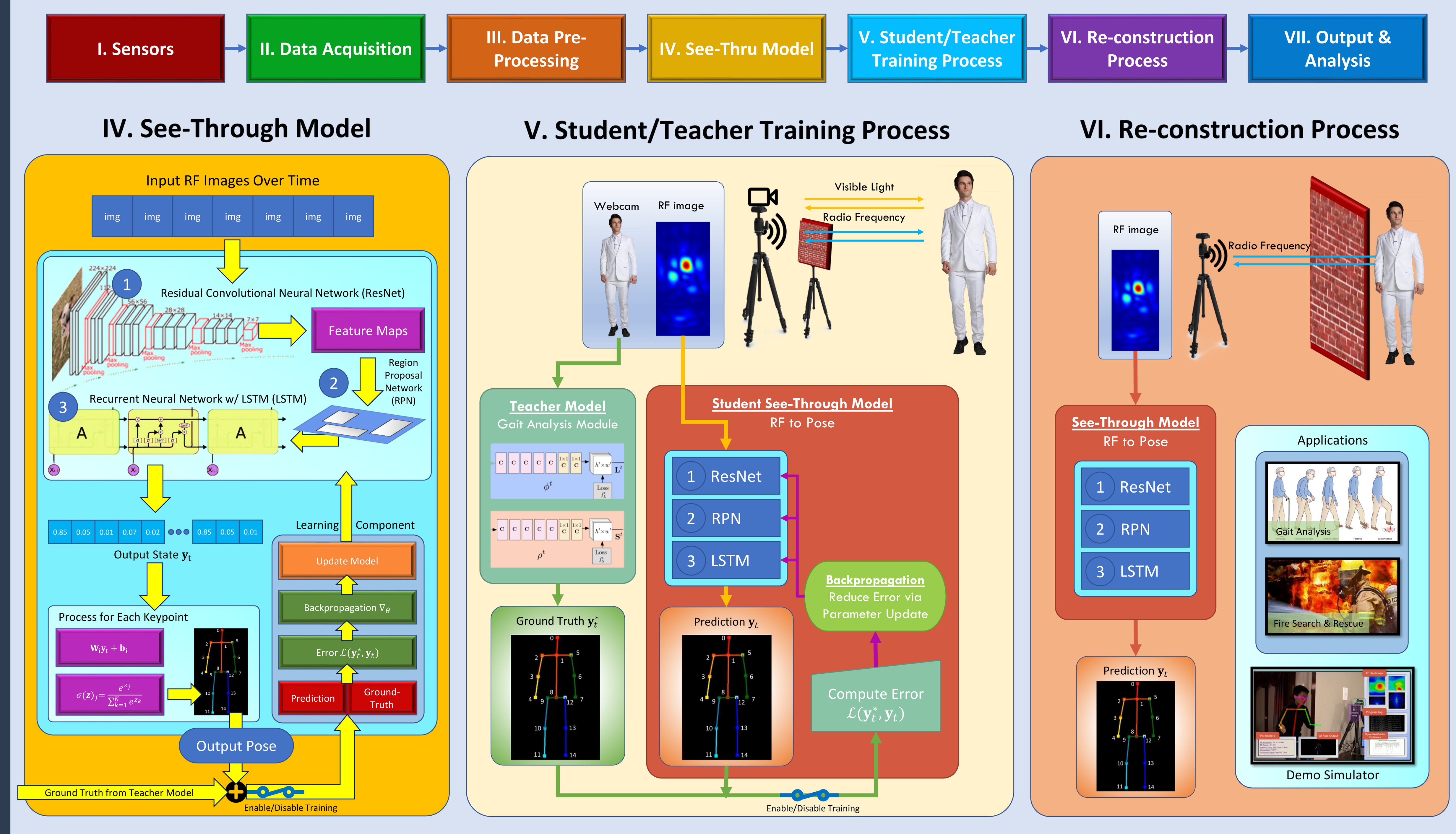
Walabot API is used to interface with RF antenna hardware, outputting a 3D matrix representing reflected signal power from each voxel in space (.csv files). The RGB camera logs video footage into .mp4 files.



First, RF data is converted into cartesian coordinates. Then, it is compressed into two 2D heatmaps, which retain 3D context, to reduce computational complexity. Dimensionality reduction is carried out as:

$$R_{horz}(x,z) = \sum_{y=y_{min}}^{y_{max}} R(x,y,z) \qquad R_{vert}(x,y) = \sum_{z=z_{min}}^{z_{max}} R(x,y,z)$$

See-Through Model Methodology Overview

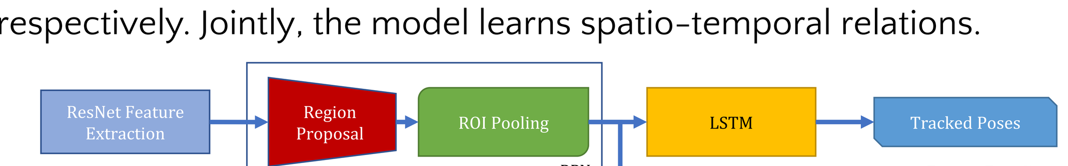


See-Thru Model (IV)

The See-Thru Model is explicitly designed to solve several challenges:

1) Low-Resolution Imaging	2) Multi-Person Support	3) Body-Part Specularity
• RF images have low intrinsic resolution • Need for translational, spatial invariance	• Multiple people must be detectable in 1 image • Movements must be tracked over time	• Body parts may be weak or missing from RF images because RF wavelength and object size are similar
Convolutional Neural Network (ResNet)	Region Proposal Network (RPN)	Recurrent Neural Net w/ Long Short-Term Memory (LSTM)

ResNet and RPN are implemented as modified DarkNet-53 and YOLO, respectively. Jointly, the model learns spatio-temporal relations.



After accumulating info over 5 RF frames, LSTM outputs logit vectors $\mathbf{p}^p$ for each person p. Then, dense layers create 15 hidden vectors $\mathbf{h}_k^p$ for each keypoint k. Finally, each keypoint is given an x/y coordinate (α/β) using two dense layers containing shared weights and biases.

$$\begin{cases} \alpha_k^p = \operatorname{argmax}\left(\mathbf{W}_\alpha \mathbf{h}_k^p + \mathbf{b}_\alpha\right) \\ \beta_k^p = \operatorname{argmax}\left(\mathbf{W}_\beta \mathbf{h}_k^p + \mathbf{b}_\beta\right) \end{cases} \forall k, p$$

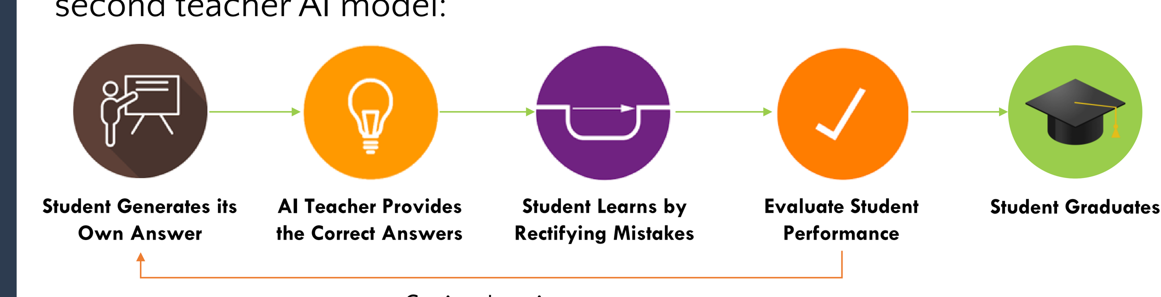
Each component of novel training objective $\mathcal{L}$ enforces:

- $\mathcal{L}_{rpn}$: 1) bounding box accuracy 2) continuity (discourages erratic-ness)
- $\mathcal{L}_{track}$: same person classified to same identity over time
- $\mathcal{L}_{cls}$: accuracy of body keypoint predictions

$$\mathcal{L} = \mathcal{L}_{\text{rpn}} + 0.75\mathcal{L}_{\text{track}} + \mathcal{L}_{\text{cls}}$$

Student/Teacher Training Process (V)

In this supervised learning task, ground-truth labels are necessary. However, it is prohibitive to manually annotate 1.5 million keypoints. A training process, inspired by the Feynman technique, is proposed to reduce the manual workload by automating label generation with a second teacher AI model:



The teacher model is implemented as a state-of-the-art Part-Affinity-Field CNN model for gait analysis, better known as OpenPose.

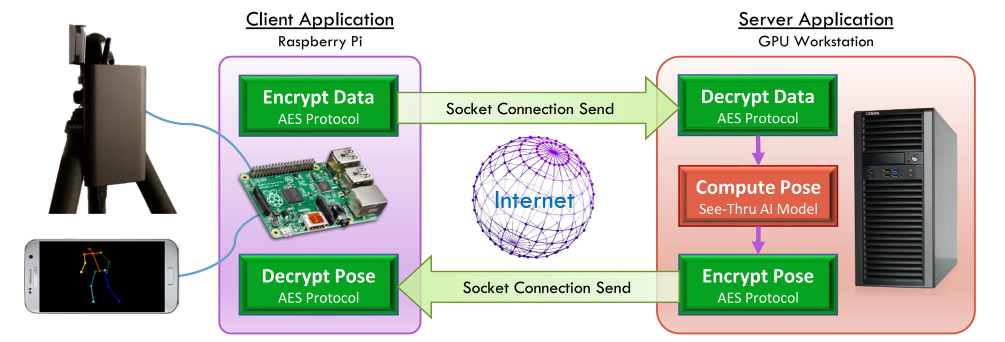
A novel training strategy is proposed. The model is explicitly trained through different materials with a supplemental stand, helping simulate wave attenuation without blocking camera. This is crucial to model performance.

Reconstruction Process (VI)

Here, train mode is switched to reconstruction mode. The teacher and RGB camera are removed. The model receives only RF signals as input.

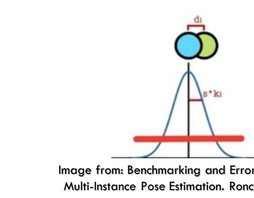
Prototype (VI)

A Raspberry Pi is linked to a GPU server over the Internet so through-wall vision can be done in real-time on a lightweight, portable setup.



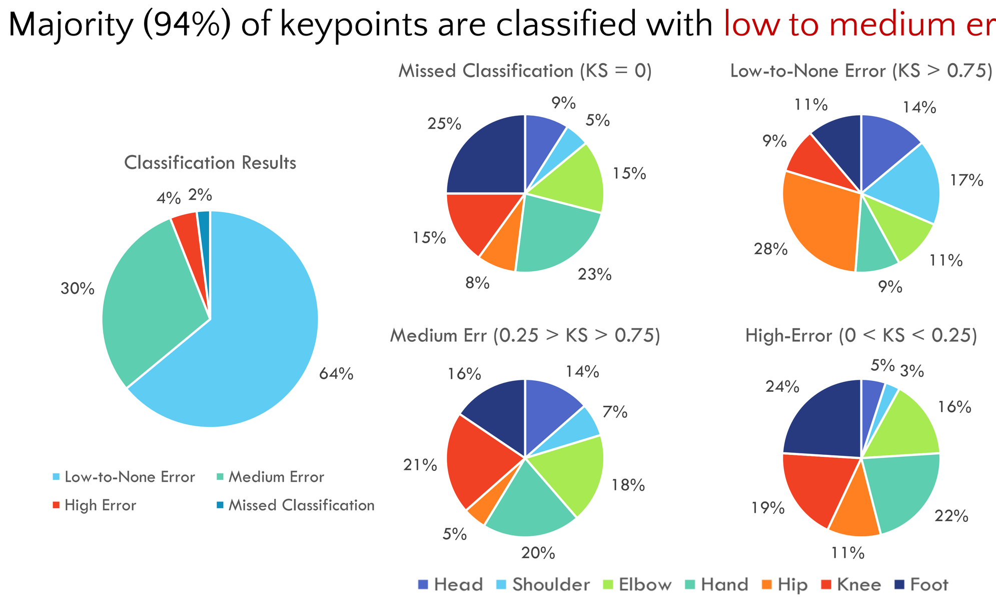
Output & Analysis (VII)

Keypoint Similarity (KS) quantifies the accuracy of each keypoint, and Object KS (OKS) is the avg KS of 15 keypoints. Where $\boldsymbol{\theta}^p$ are classified keypoints of p, $\boldsymbol{\theta}^{p*}$ are ground truth locations, C_i is the prediction's confidence level, and s and k_i are used to scale the distribution.

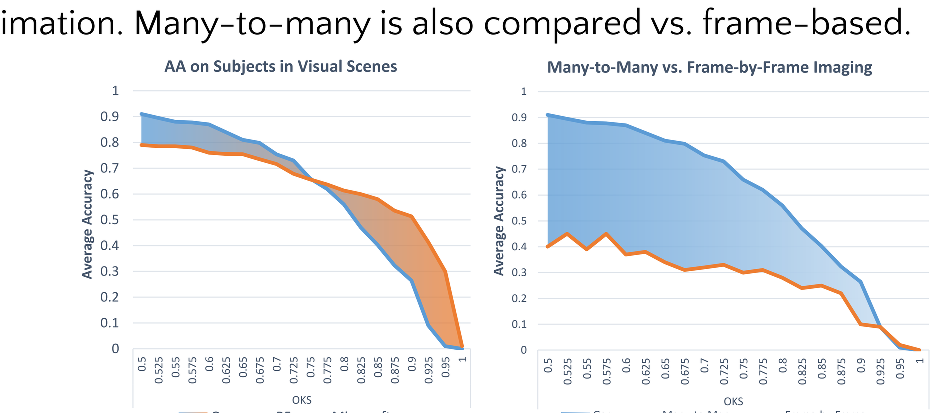


$$\begin{cases} OKS(\boldsymbol{\theta}^{p*}, \boldsymbol{\theta}^p) = \dfrac{\sum_i KS(\boldsymbol{\theta}_i^{p*}, \boldsymbol{\theta}_i^p)\delta(C_i > 0)}{\sum_i \delta(C_i > 0)} \\ KS(\boldsymbol{\theta}_i^{p*}, \boldsymbol{\theta}_i^p) = e^{-\frac{\|\boldsymbol{\theta}_i^{p*} - \boldsymbol{\theta}_i^p\|_2^2}{2s^2k_i^2}} \end{cases}$$

Majority (94%) of keypoints are classified with low to medium error.



Average Accuracy is the proportion of keypoints that have a $KS \geq$ threshold T_{KS}. RF model is compared to Microsoft's Human-Pose-Estimation. Many-to-many is also compared vs. frame-based.



RF outperforms Microsoft for lower KS thresholds. This could be attributed to the many-to-many model and the novel cost function. Additionally, many-to-many outperforms frame-by-frame.

Conclusions

This project establishes a new approach to a problem that has been investigated for over 10 years: visualizing human figures even behind obstruction. Using state-of-the-art deep learning, high performance is achieved, and especially with the novel many-to-many architecture, this technology is the closest it's ever been to practical usage.

Discussions

There are several killer applications worth mentioning:

1. RF can detect victims even in fire rescue situations, where neither visible light nor thermal infrared imaging can traverse fire & smoke pollution.
2. Improve elderly citizen monitoring with fall detection/gait disorder diagnosis, while preserving patient privacy, eliminating dead zones, and retaining detail.
3. Help military and law-enforcement leaders generate strategically sound decisions by serving presence and movement information of targets.

Current limitation cases include when people obstruct others and over-population of a scene (>3). Stereoscopic RF vision could be used to address shortcomings in the future.